GREAT LEARNING

PGP-DSBA Online

Feb'21 Batch

Date: 16/05/2021

References: ANOVA.pdf, Monograph-ANOVA.pdf

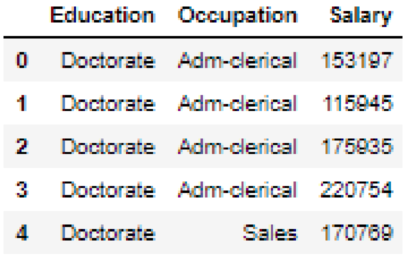
Submitted By: Prachi Gupta

ANALYSIS OF VARIANCE(ANOVA) OF SALARY DATA

Problem Statement:

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

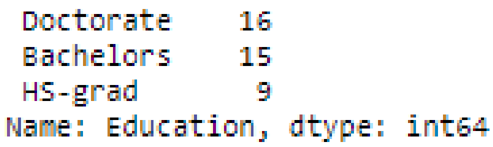
Sample of Dataset: Shown in **Figure: 1**



	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769

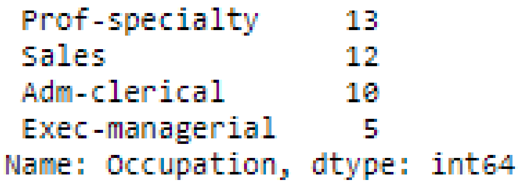
Figure: 1

Dataset has 3 variables, Education, Occupation and Salary.There are 3 levels of Education(**Figure: 2)** & 4 levels of Occupation(**Figure: 3)**



```
 Doctorate    16
 Bachelors    15
 HS-grad       9
Name: Education, dtype: int64
```

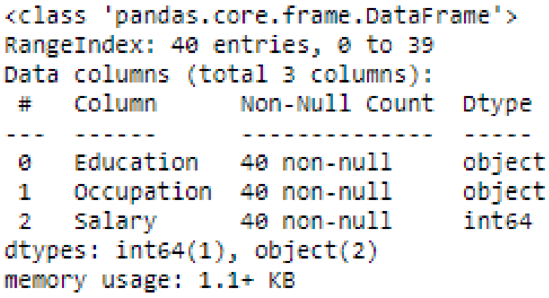
Figure: 2



```
 Prof-specialty     13
 Sales              12
 Adm-clerical       10
 Exec-managerial     5
Name: Occupation, dtype: int64
```

Figure: 3

'Salary' is a dependent , continuous variable of int type
'Education' & 'Occupation' are categorical variables of object types.
None of the columns have any null value.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Education   40 non-null     object
 1   Occupation  40 non-null     object
 2   Salary      40 non-null     int64 
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

Figure: 4

Assumptions made:

1. Salary data was drawn from a population having normal distribution.
2. All input samples are from populations with equal variances.
3. The sample is a random sample ,i.e., observations shown in sample are collected independently of each other.

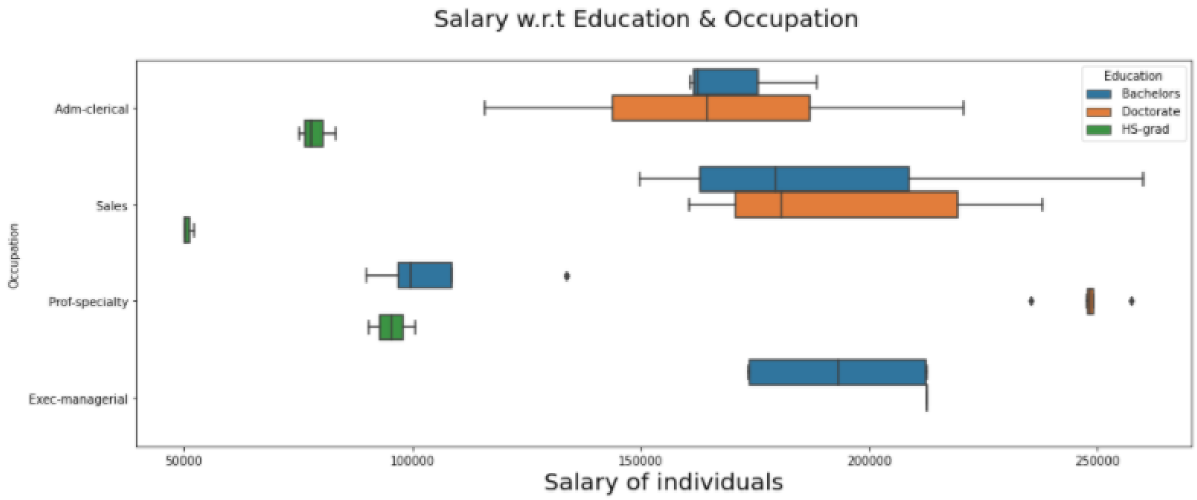


Figure: 5 (Boxplot-Visual comparison of group means)

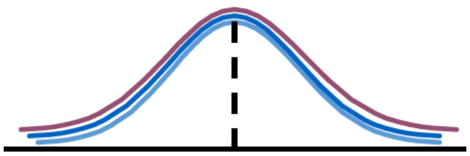
Problem 1A:

1. **State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.**

The objective is to determine whether Salaries of individuals depend on their educational qualification or occupation or both.

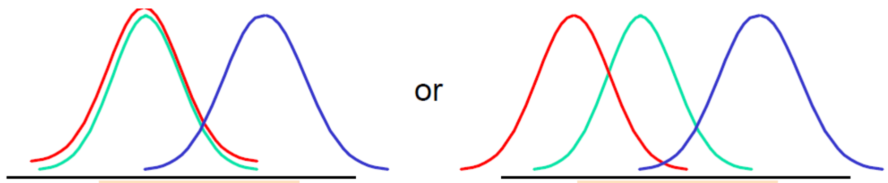
a) The Hypothesis made for the One Way ANOVA for **Education** are:

Null Hypothesis **Ho:** The mean salary earned by an individual is same with different categories of Education.



$\mu_1 = \mu_2 = \mu_3$

Alternative Hypothesis **Ha:** The mean salary earned by an individual is different in at-least one category of Education qualification.



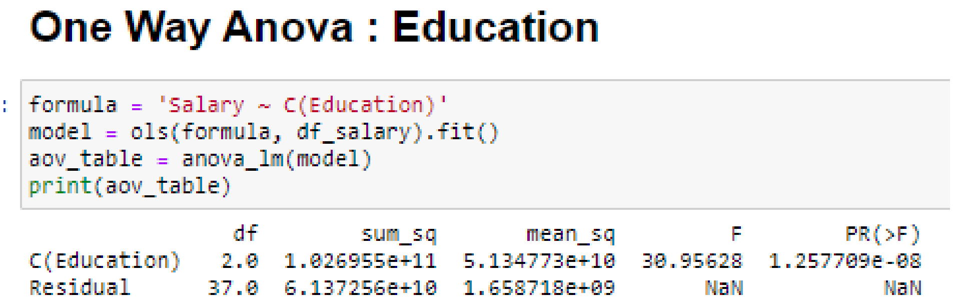
$\mu_1 = \mu_2 \neq \mu_3$ $\mu_1 \neq \mu_2 \neq \mu_3$

b) The Hypothesis made for the One Way ANOVA for **Occupation** are:

Null Hypothesis **Ho:** The mean salary earned by an individual is same with different categories of Occupation.(No Factor Effect exists)

Alternative Hypothesis **Ha:** The mean salary earned by an individual is different in at-least one category of Occupation.(Factor Effect is present)

2. **Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.**



One Way Anova : Education

```
formula = 'Salary ~ C(Education)'
model = ols(formula, df_salary).fit()
aov_table = anova_lm(model)
print(aov_table)
```

```
                df        sum_sq       mean_sq          F        PR(>F)
C(Education)   2.0  1.026955e+11  5.134773e+10  30.95628  1.257709e-08
Residual      37.0  6.137256e+10  1.658718e+09       NaN           NaN
```

Figure: 6

The calculated f-statistic value: 30.95628008792558
p-value: 1.2577090926629002e-08

Conclusion: We see in **Figure: 6** that the corresponding p-value is less than the **significance level** (0.05). Thus, we **reject the Null Hypothesis** and conclude that the mean salary earned by an individual is different in at-least one category of Education qualification.
Below point plot (**Figure: 7)** also shows the increase in salaries as the Education qualification increases.

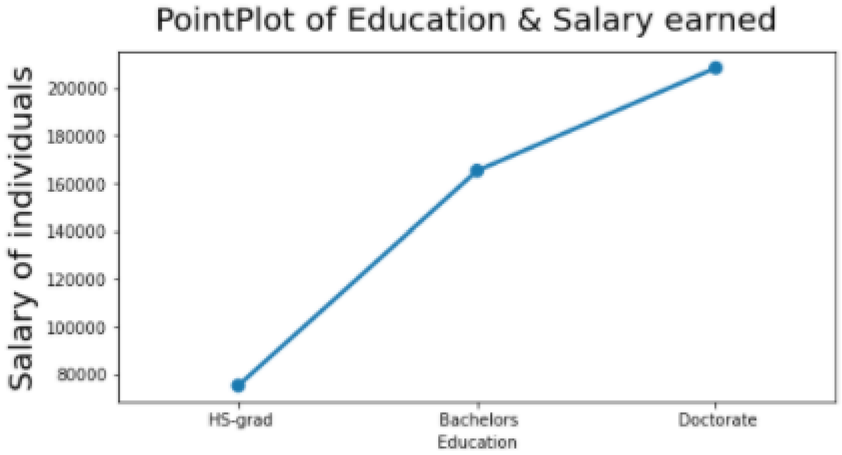
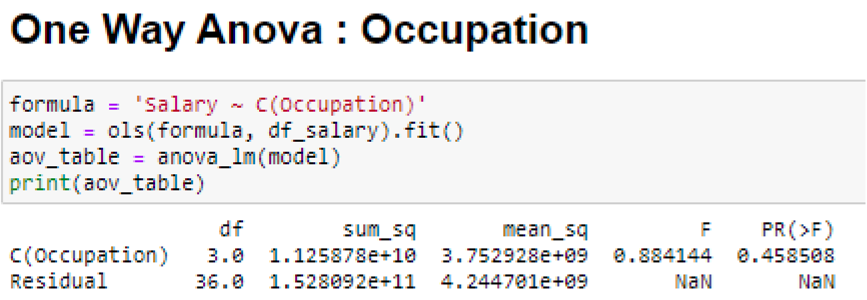


Figure: 7

3. **Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.**



One Way Anova : Occupation

```
formula = 'Salary ~ C(Occupation)'
model = ols(formula, df_salary).fit()
aov_table = anova_lm(model)
print(aov_table)
```

```
                  df        sum_sq       mean_sq         F    PR(>F)
C(Occupation)    3.0  1.125878e+10  3.752928e+09  0.884144  0.458508
Residual        36.0  1.528092e+11  4.244701e+09       NaN       NaN
```

Figure: 8

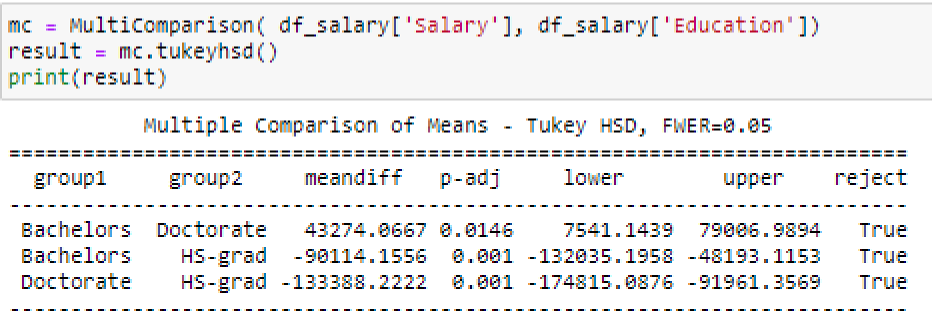
The calculated f-statistic value: 0.8841441289216039
p-value: 0.4585078266495116

Conclusion: Since the p value is more than the **significance level** (0.05) in **Figure: 8**, we **fail to reject the null hypothesis** and conclude that the mean salary earned by an individual is same with different categories of Occupation.

4. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result

Since the null hypothesis is rejected in (2), we perform Tukey HSD (Honestly Significant Difference) Test to determine differences of means class - wise. This provides us the differences of means of 2 levels, taken at a time, based on which it concludes if Null Hypothesis should be rejected/accepted for that combination. Here,

Doctorate - Bachelors = 43274.0667
HS-grad - Bachelors = -90114.1556
HS-grad - Doctorate = -133388.2222



```
mc = MultiComparison( df_salary['Salary'], df_salary['Education'])
result = mc.tukeyhsd()
print(result)
```

```
           Multiple Comparison of Means - Tukey HSD, FWER=0.05
=========================================================================
  group1     group2     meandiff   p-adj     lower         upper    reject
-------------------------------------------------------------------------
 Bachelors  Doctorate   43274.0667 0.0146    7541.1439  79006.9894   True
 Bachelors    HS-grad  -90114.1556  0.001 -132035.1958 -48193.1153   True
 Doctorate    HS-grad -133388.2222  0.001 -174815.0876 -91961.3569   True
-------------------------------------------------------------------------
```

Figure: 9

Hence, Doctorate > Bachelors > HS-grad

The result from **Figure: 9** reveals that the means of ‘Bachelors’ category of ‘Education’ is significantly different from the means of HS-grad category of ‘Education’.

Problem 1B:

1. **What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.**

The interaction between two factors/treatments is used to understand how the relationship between one categorical factor and a continuous response depends on the value of second categorical factor present in the dataset.

Below is the interaction plot (**Figure: 10)** showing association among both treatments. It displays the means for levels of one factor on x-axis and a separate line for each level of another factor. As the plots are non-parallel, this implies that an interaction exists between them.

Observation:
This (**Figure: 10)** indicates that the relationship between 'Occupation' and 'Salary' depends on the category of 'Education' given.

For an 'Education' level of '**HS-grad**', an individual with Occupation as '**Sales'** is associated with **lowest** mean 'Salary' value, and a **'Prof-specialty'** one is associated with **highest** mean 'Salary' value.

For an 'Education' level of '**Bachelors'**, an individual with Occupation as **'Prof-specialty'** is associated with **lowest** mean 'Salary' value, and a 'Sales' & 'Exec-managerial' are associated with **highest** mean 'Salary' value.

For an 'Education' level of '**Doctorate**, an individual with Occupation as **'Adm-Clerical'** is associated with **lowest** mean 'Salary' value, and a **'Prof-specialty'** one is associated with **highest** mean 'Salary' value.

The interaction term is present across all the levels of 'Education' with the categories of 'Occupation' here. That's why, an individual with highest/lowest value for an educational level doesn't follow the same trend for another level.

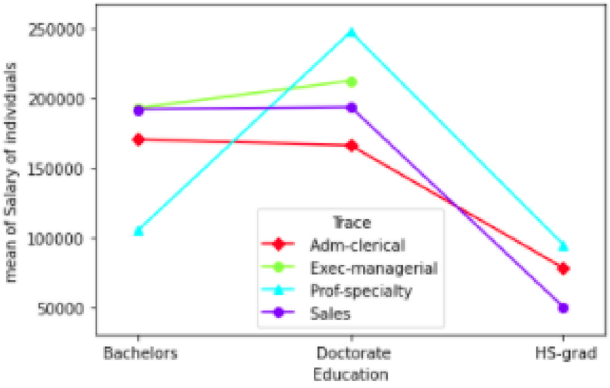


Figure: 10

2. **Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?**

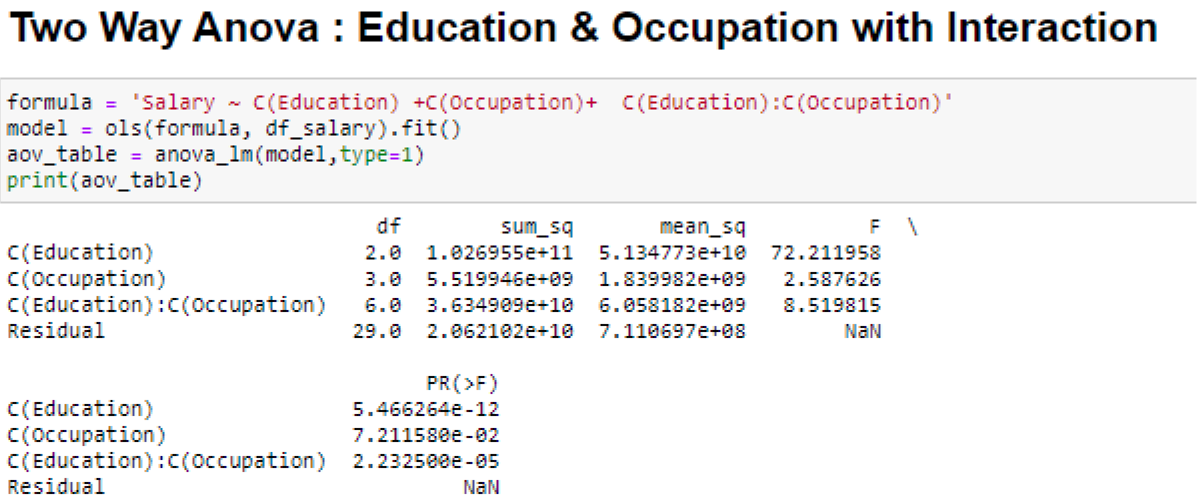
The objective is to determine whether Salaries of individuals depend on both their educational qualification and occupation , as well as on the interaction of both.
The Hypothesis made for the two-way ANOVA are:
Null Hypothesis **Ho:**

1. The mean salary earned by an individual is same with different categories/ levels of Education.
2. The mean salary earned by an individual is same with different categories/ levels of Occupation.

3. The mean salary earned by an individual is same with both factors considered together.There is no interaction parameter between the two factors Education : Occupation.

Alternate Hypothesis **Ha:**

1. The mean salary earned by an individual is different with different categories/ levels of Education.
2. The mean salary earned by an individual is different with different categories/ levels of Occupation.
3. The mean salary earned by an individual is different with both factors considered together.There is an interaction between the two factors Education & Occupation i.e., the effect that one factor has on the other.



Two Way Anova : Education & Occupation with Interaction

```
formula = 'Salary ~ C(Education) +C(Occupation)+  C(Education):C(Occupation)'
model = ols(formula, df_salary).fit()
aov_table = anova_lm(model,type=1)
print(aov_table)
```

```
                              df        sum_sq       mean_sq          F  \
C(Education)                 2.0  1.026955e+11  5.134773e+10  72.211958   
C(Occupation)                3.0  5.519946e+09  1.839982e+09   2.587626   
C(Education):C(Occupation)   6.0  3.634909e+10  6.058182e+09   8.519815   
Residual                    29.0  2.062102e+10  7.110697e+08        NaN   

                                  PR(>F)  
C(Education)                5.466264e-12  
C(Occupation)               7.211580e-02  
C(Education):C(Occupation)  2.232500e-05  
Residual                             NaN  
```

Figure: 11

Due to the inclusion of the interaction effect term, we can see a slight decrease in the p-values of the first two treatments as compared to the Two-Way ANOVA without the interaction effect terms. And we see that the p-value of the interaction effect term of 'Education' and 'Occupation' is less than 0.05, which suggests that the Null Hypothesis is rejected in this case (**Figure: 11)**.

Conclusion:

There is no Influence of the change of levels of the factor ‘Occupation’ in ‘Salary’, as p-value > 0.05. So, we fail to reject Ho.
Hence, all the means of ‘Salary’ for the 4 categories of ‘Occupation’ are the same.
With different categories of ‘Occupation’, mean ‘Salary’ of an individual does not change.

There is an Influence of the factor ‘Education’ in ‘Salary’, as p-value < 0.05. So, Ho is rejected
Hence, all the means of ‘Salary’ for the 3 categories of ‘Education’ are different.
With different categories of ‘Education’, mean ‘Salary’ of an individual changes.

There is an Influence of the Interaction parameter of factors, ‘Education’: ‘Occupation’ in ‘Salary’, as p-value < 0.05. So, Ho is rejected
Hence, the means of the ‘Salary’ of individuals are different when we consider both the factors, ‘Education’ & ‘Occupation’ together.

3. Explain the business implications of performing ANOVA for this particular case study.

Business Implications derived from this case study are as follows:

- Mean ‘Salary’ of an individual increases as the ‘Education’ level increases from ‘HS-grad’ to ‘Doctorate’.
- There is no change in mean salary across the 4 levels of ‘Occupation’.
- Since there is an interaction between ‘Education’ & ‘Occupation’ levels, this implies that for a particular Occupation level, mean ‘Salary’ increases along with the increase in ‘education’ level.

- It is highly recommended for any particular category of ‘Occupation’ that the ‘Education’ qualification of the individual should be kept at the lowest possible level. For example, For the occupation level of ‘Exec-managerial’, the ‘Education’ level can be kept as ‘Bachelors’ instead of ‘Doctorate’